Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*

Stanley A. Sawyer*, John Parsch†, Zhi Zhang†, and Daniel L. Hartl‡§

*Department of Mathematics, Washington University, St. Louis, MO 63130; †Section of Evolutionary Biology, Department of Biology II, University of Munich, 82152 Munich, Germany; and ‡Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected on May 3, 2005.

Contributed by Daniel L. Hartl, February 21, 2007 (sent for review February 7, 2007)

We have estimated the selective effects of amino acid replacements in natural populations by comparing levels of polymorphism in 91 genes in African populations of *Drosophila melanogaster* with their divergence from *Drosophila simulans*. The genes include about equal numbers whose level of expression in adults is greater in males, greater in females, or approximately equal in the sexes. Markov chain Monte Carlo methods were used to sample key parameters in the stationary distribution of polymorphism and divergence in a model in which the selective effect of each nonsynonymous mutation is regarded as a random sample from some underlying normal distribution whose mean may differ from one gene to the next. Our analysis suggests that ≈95% of all nonsynonymous mutations that could contribute to polymorphism or divergence are deleterious, and that the average proportion of deleterious amino acid polymorphisms in samples is ≈70%. On the other hand, ≈95% of fixed differences between species are positively selected, although the scaled selection coefficient ($N_e s$) is very small. We estimate that ≈46% of amino acid replacements have $N_e s < 2$, ≈84% have $N_e s < 4$, and ≈99% have $N_e s < 7$. Although positive selection among amino acid differences between species seems pervasive, most of the selective effects could be regarded as nearly neutral. There are significant differences in selection between sex-biased and unbiased genes, which relate primarily to the mean of the distributions of mutational effects and the fraction of slightly deleterious and weakly beneficial mutations that are fixed.

McDonald–Kreitman test | polymorphism and divergence | protein evolution

Synonymy in the genetic code results in a natural periodicity in which the third nucleotide of many codons is only weakly constrained because any of two or more nucleotides at this position specify the same amino acid in the polypeptide chain. Fourfold degenerate codons allow any nucleotide at the third position, whereas twofold degenerate codons treat either both pyrimidine nucleotides or both purine nucleotides as synonymous. Of the 20 common amino acids, the codons for 12 are twofold degenerate at the third position, 1 is threefold degenerate (isoleucine, which allows U, C, or A at the third position), and 8 are fourfold degenerate. (In this tabulation, leucine, serine, and arginine are each counted twice because each is specified by six codons.) In a typical coding sequence with a GC content of 50% the average codon degeneracy is 3.

The high level of synonymy in the genetic code is a boon to population genomics, because the synonymous sites in a coding sequence serve as a sort of internal control for historical and demographic factors affecting a population, relatively free of selective constraint. Because nonsynonymous sites in the same coding sequence share the same history and demography as the synonymous sites, but may be subject to greater selective constraints or even positive selection, comparisons between nonsynonymous sites and synonymous sites can potentially reveal the magnitude and direction of selection pressures operating on the nonsynonymous sites.

An early application of this approach compared the frequency spectrum of polymorphic nonsynonymous sites with that of synonymous sites among sequences encoding 6-phosphogluconate dehydrogenase in a sample of the enteric bacterium *Escherichia coli* (1). An excess of low-frequency nonsynonymous polymophisms suggested that most amino acid polymorphisms in this enzyme are very slightly deleterious, with a selection coefficient on the order of 6–26 times the reciprocal of the effective population size. No more than half of all amino acid polymorphisms in the enzyme could be considered as selectively neutral.

An important extension of this approach came from McDonald and Kreitman (2), who compared polymorphisms within species to divergence between species. This approach avoided any need to estimate the allele-frequency spectrum of polymorphisms, while taking advantage of evolutionary changes through time. First applied to the *Adh* gene encoding alcohol dehydrogenase in three species of the *Drosophila melanogaster* species subgroup, the approach yielded evidence that a significant proportion of amino acid replacements between species are driven by positive selection. Explicit expressions for the expected values in comparisons of polymorphism and divergence were soon developed based on a sampling theory for the independent infinite-sites model with selection (3). Application of this theory to the *Drosophila Adh* data again suggested small selection coefficients, on the order of five times the reciprocal of the effective population size, and that the number of amino acids in the enzyme that are susceptible to favorable mutation at any one time ranges from 2 to 23.

One limitation of the McDonald–Kreitman test is that, for the sample sizes typically available, the statistical test for homogeneity in a 2×2 table is relatively lacking in power. Another limitation is that such data often include one or more cells whose entry is 0. Thus there has been an effort to examine polymorphism-divergence data across multiple genes to estimate α, defined as the fraction of amino acid fixations driven by positive selection (4, 5). Maximum-likelihood approaches yield estimates of α of $25\% \pm 20\%$ across several species of *Drosophila* (6, 7). This approach assumes that harmful mutations are so drastically deleterious, and beneficial mutations so strongly favored, that their fate is settled so rapidly by selection that they cannot contribute significantly to the level of amino acid polymorphism. Considerable evidence suggests that this assumption is not correct (1, 5, 8–10). To the extent that mildly deleterious and mildly favorable nonsynonymous substitutions contribute to amino acid polymorphisms, the estimate of α is biased down-

Author contributions: S.A.S. and J.P. contributed equally to this work; S.A.S., J.P., and D.L.H. designed research; J.P. and Z.Z. performed research; S.A.S. contributed new reagents/analytic tools; S.A.S. and D.L.H. analyzed data; and S.A.S., J.P., and D.L.H. wrote the paper.

The authors declare no conflict of interest.

§To whom correspondence should be addressed. E-mail: dhartl@oeb.harvard.edu.

ward. The assumption of fluctuating selection leads to somewhat higher estimates (11).

Quite another approach to the analysis of polymorphism and divergence makes use of population genetics theory (3) to estimate the values of the parameters governing mutation, selection, and random genetic drift at independent nucleotide sites (12). The intuitive appeal of this approach is that it avoids the artificial dichotomy between what is selectively neutral and what is not, but rather focuses on the actual estimates of the selection coefficients that emerge from the analysis. In this model, the expected value of each cell in a McDonald–Kreitman table can be shown to be an independent Poisson random variable (3), and the parameters governing mutation, selection, random genetic drift, and time since species divergence can be estimated by Markov chain Monte Carlo simulation using a hierarchical Bayesian model (12). In the original formulation, each nonsynonymous substitution likely to contribute to polymorphism or divergence in a particular gene is assumed to have the same selective effect, but these values can differ from one gene to the next. The selective effect is scaled according to the diploid effective population number, which is to say that it is estimated as some multiple of $N_e s$, where s is the conventional selection coefficient and N_e is the diploid effective population size. This approach is reliable provided that the species being compared are sufficiently closely related that multiple nucleotide substitutions at the same site, or synonymous sites mutating to nonsynomous sites or vice versa, can be ignored (13).

The assumption that each nonsynonymous substitution in a gene has the same selective effect is obviously artificial, but it served the original purpose of estimating the distribution of the scaled selection coefficient among genes (12). A more sophisticated and biologically realistic model was introduced by Sawyer *et al.* (9). In this model, the selective effect of each nonsynonymous mutation likely to contribute to polymorphism or divergence is regarded as a random sample from some underlying normal distribution whose mean but not variance may differ from one gene to the next. The spirit of the model is analogous to that of analysis of variance, in which different "treatments" (in this case, genes) have different "effects" (in this case, mean selective effects). The assumption that the underlying distributions are Gaussian is natural in a continuous-time model of selection (14) given the implications of the Central Limit Theorem, but plausible alternatives should also eventually be considered.

Changes in demographics can confound the interpretation of polymorphism and divergence (2, 5, 15). For example, a rapid dramatic increase in the effective population number will result in the selective elimination of some deleterious nonsynonymous polymorphisms that might previously have remained polymorphic, thereby reducing the nonsynonymous polymorphisms without affecting nonsynonymous divergence. Demographics need to be considered for the sibling species *D. melanogaster* and *Drosophila simulans*, which appear to have expanded their range out of Africa ≈10,000–15,000 years ago (16, 17), probably with an accompanying a population bottleneck followed by an expansion (18).

Hence, for *Drosophila* the ideal polymorphism data would seem to be that derived from African populations. As it happens, Pröschel *et al.* (19) have recently acquired such data for a large set of genes. These data afford a valuable opportunity to apply the Sawyer model (9) to estimate values of great interest in population genomics, including the fraction of amino acid polymorphisms that are deleterious, the fraction of amino acid differences between related species that are nearly neutral or positively selected, and the distribution of selection coefficients among new mutations likely to become polymorphic or among mutations that are fixed. In this article we present the results of the analysis. The principal inferences are that the majority of amino acid polymorphisms within *Drosophila* species are mildly deleterious but that a large fraction of amino acid differences between species are driven by positive selection. However, the magnitude of selection that needs to be postulated to explain the data is extremely small, usually >2 but <10 times the reciprocal of the effective population size. These results are predicated on the assumption that most synonymous polymorphisms and fixed differences are selectively neutral or nearly neutral, and so they pertain only to amino acid substitutions and not to nucleotide substitutions in noncoding DNA.

Results

Data. The Pröschel *et al.* (19) data consist of the coding sequences of up to 12 alleles of each of 91 genes in samples of *D. melanogaster* derived from Lake Kariba, Zimbabwe (20). Among these genes are 33 that are male-biased in their expression, 28 that are female-biased, and 30 that are equally expressed in the sexes (unbiased). Sex-biased expression means at least a 2-fold expression difference between males and females (or between testes and ovaries) as estimated in microarray experiments (21–23), and unbiased expression means a ratio of expression in the range 0.75–1.25 (19). These polymorphism data were compared with divergence from a highly inbred line of *D. simulans* from Chapel Hill, North Carolina (24) to estimate α, the proportion of amino acid replacements subject to positive selection (6), and the distribution of scaled selection coefficients across genes (12), to test for differential selection between sex-biased and unbiased genes (19). Here, we describe and apply a model that relaxes the assumption that the selection coefficient is identical for all amino acid substitutions in each gene. This model allows us to estimate quantitatively the distribution of selection coefficients within and among loci and the fraction of amino acid replacements between species that are selectively neutral or nearly neutral.

Random-Effects Model of Selection. For the sake of generality, consider a set of aligned coding sequences without gaps representing m alleles sampled from one species and n alleles sampled from the orthologous gene in a related species. The species are assumed to be sufficiently closely related that multiple substitutions of the same nucleotide are unlikely. We shall disregard all codons that are monomorphic across both samples and classify the others into one of four categories: synonymous divergence (both samples monomophic but differ in a synonymous codon), synonymous polymorphic (one or both samples polymorphic for a synonymous codon), replacement divergent (both samples monomophic but differ in a nonsynonymous codon), or replacement polymorphic (one or both samples polymorphic for a nonsynonymous codon). These four counts form a 2×2 McDonald–Kreitman table (2) for the alleles of any one locus, and for any set of k loci they form a group of k such tables.

We assume that all synonymous substitutions are selectively neutral or nearly neutral but that nonsynonymous substitutions are each potentially subject to selection. Our objective is to estimate the distribution of selection coefficients of the nonsynonymous substitutions at each locus. We assume a population of constant and finite size reproducing continuously in time, so that the appropriate measure of relative fitness is the malthusian parameter defined as the natural logarithm of the Darwinian fitness (14).

Suppose that at the ith locus the distribution of selection coefficients is normal with mean γ_i and variance σ_w^2, where the within-locus variance σ_w^2 is the same for each locus. Symbolically, we can write the distribution of selection coefficients for new mutations at the ith locus as

$$\gamma \sim \gamma_i + \sigma_w N(0, 1), \qquad [1]$$

where $N(0, 1)$ is a random sample from a standard normal distribution. Across loci, the selection coefficients have a normal distribution with variance $\sigma_b^2 + \sigma_w^2$, where σ_b^2 is the between-locus variance and is equal to the variance among the γ_i values. These assumptions imply that the selection coefficients for two new mutations at the same locus are normally distributed with variance $\sigma_b^2 + \sigma_w^2$ and covariance σ_b^2. The γ values are scaled according to two times the diploid effective population size, hence $\gamma/2 = N_e s$, where N_e is the diploid effective population number and s is the conventional selection coefficient.

At equilibrium in a large population, each nonsynonymous substitution that is polymorphic at the ith locus can be described by a pair (y, p) where y is drawn from a normal distribution with mean γ_i (the mean scaled selection coefficient) and p is the proportion of the population that carries the nonancestral nucleotide at the site. The pairs (y, p) are the points of a 2D Poisson random field on $(-\infty, +\infty) \times (0, 1)$. Likewise, if we define T as the time since divergence of the two species and assume that T is large enough for mutation-selection-drift equilibrium to have been attained, the selection coefficients of nonsynonymous fixed differences between the species form a Poisson random field on $(-\infty, +\infty)$. Similar considerations apply to polymorphic and divergent synonymous sites, except that the selection coefficients are all set to 0.

For nonsynonymous polymorphic sites, the mean density of the Poisson random field is given by

$$\theta_{r,i}\phi(y, \gamma_i, \sigma_w)\frac{1 - \exp[-2y(1-p)]}{1 - \exp(-2y)}\frac{1}{p(1-p)}\,dy\,dp, \qquad [2]$$

where $\theta_{r,i}$ is the rate of mutation to nonsynonymous nucleotides that have a reasonable chance of becoming polymorphic or fixed. The magnitude of $\theta_{r,i}$ is scaled by $4N_e$, where N_e is the diploid effective population size. We stipulate that the only mutations under consideration are those that have a reasonable chance of becoming polymorphic or fixed because the polymorphism and divergence of samples are uninformative about mutations whose deleterious effects are so severe that they are very unlikely ever to be present in a sample. In Eq. **2**, $\phi(y, \gamma_i, \sigma_w)$ is a normal probability density function for a random variable y with mean γ_i and variance σ_w^2.

For nonsynonymous fixed differences, the mean density of the Poisson random field is given by

$$T\theta_{r,i}\phi(y, \gamma_i, \sigma_w)\frac{2y}{1 - \exp(-2y)}\,dy, \qquad [3]$$

where T is the time in generations since species divergence, scaled by two times the diploid effective population size. Eqs. **2** and **3** are extensions of Wright's formulas (25), for which a different proof was given in Sawyer and Hartl (3) as part of a derivation of formulas for Poisson random fields.

The corresponding mean densities of the Poisson random fields for polymorphism and divergence of synonymous sites are, respectively:

$$\theta_{s,i}\frac{1}{p}\,dp \qquad [4]$$

$$T\theta_{s,i}, \qquad [5]$$

where $\theta_{s,i}$ is the rate of mutation to synonymous nucleotides, scaled by four times the diploid effective population size. Because all synonymous mutations are assumed to be selectively neutral and so have a reasonable chance of becoming polymorphic, $\theta_{s,i}$ includes all synonymous mutations.

Eqs. **2**–**5** imply that each of the cells in a McDonald–Kreitman table for one locus has a count whose magnitude is distributed as a Poisson distribution whose mean is shown in Table 1, where $a_1(m) = 1 + 1/2 + 1/3 + \ldots + 1/(m-1)$ and

Table 1. Expected counts for polymorphism and divergence

	Divergence	Polymorphism
Synonymous	$\theta_{s,i}\left(T + \frac{1}{m} + \frac{1}{n}\right)$	$\theta_{s,i}[a_1(m) + a_1(n)]$
Replacement	$\theta_{r,i}[G_0(\gamma_i, \sigma_w)T + G_1(\gamma_i, \sigma_w, m) + G_1(\gamma_i, \sigma_w, n)]$	$\theta_{r,i}[G_2(\gamma_i, \sigma_w, m) + G_2(\gamma_i, \sigma_w, n)]$

$$G_0(\gamma_i, \sigma_w) = \int_{-\infty}^{\infty}\phi(y, \gamma_i, \sigma_w)\frac{2y}{1 - e^{-2y}}\,dy \qquad [6]$$

$$G_1(\gamma_i, \sigma_w, m) = \int_{-\infty}^{\infty}\phi(y, \gamma_i, \sigma_w)\frac{2y}{1 - e^{-2y}}\left[\int_0^1 (1-p)^{m-1}\frac{1 - e^{-2yp}}{2yp}\,dp\right]dy \qquad [7]$$

$$G_2(\gamma_i, \sigma_w, m) = \int_{-\infty}^{\infty}\phi(y, \gamma_i, \sigma_w)\frac{2y}{1 - e^{-2y}}\left[\int_0^1 \frac{1 - p^m - (1-p)^m}{1-p}\frac{1 - e^{-2yp}}{2yp}\,dp\right]dy. \qquad [8]$$

In Table 1, the term involving G_0 counts the number of nonsynonymous substitutions that are already fixed between species, whereas the terms involving G_1 count the number of nonsynonymous substitutions that are fixed differences in the sample but polymorphic in one or both populations. Similarly, the term involving G_2 counts the number of nonsynonymous substitutions that are polymorphic in the sample and in the population.

For polymorphism-divergence data across a set of k loci, each of the k loci has three locus-specific parameters ($\theta_{r,i}$, $\theta_{s,i}$, and γ_i). The model also has four global parameters (T, σ_w, σ_b, and μ), where μ is the average selection coefficient of new nonsynonymous mutations across loci. These $3k + 4 = 277$ parameters in the Pröschel *et al.* data (19) were estimated by means of sampling from Monte Carlo Markov chains whose stationary distributions simulate those of the mutation-selection-drift process (26, 27). Details are described in *Methods*.

Distribution of Selection Coefficients Among New Mutations. In the model, values of the scaled selection coefficient at the ith locus are assumed to be drawn from a normal distribution $N(\gamma_i, \sigma_w)$ with mean γ_i and standard deviation σ_w. Across loci, the γ_i are distributed as $N(\mu, \sigma_b)$. Estimates of these parameters were obtained from the average across 200,000 samples from each of 10 subchains and are presented here as multiples of the diploid population size N_e.

Fig. 1 shows the distributions of the estimated scaled selection coefficients ($\gamma_i/2 = N_e s$) for genes whose expression in mature flies is male-biased (red), female-biased (green), or unbiased (blue). Scaled to the diploid population size, the global parameter estimates are $\mu = -5.7 \pm 15.5$ and $\sigma_b = 2.1 \pm 2.2$, and within each locus the standard deviation is estimated as $\sigma_w = 3.5 \pm 5.7$. The inferred distributions in Fig. 1 support the commonplace belief that most nonsynonymous mutations are deleterious. The nonsynonymous mutations included in Fig. 1 are only a subset of

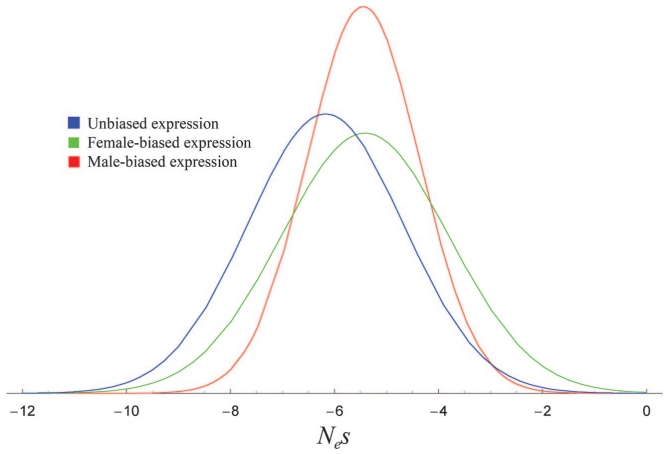

Fig. 1. Inferred distribution of scaled selection coefficients $N_e s$ among new nonsynonymous mutations that could plausibly become polymorphic or fixed, where N_e is the diploid effective population number and s is the conventional selection coefficient. $N_e s$ corresponds to the parameter $\gamma/2$ in Eq. **1**. The mutational distributions exclude all mutations that are lethal or sterile and those with selective effects that are so deleterious as to preclude their becoming polymorphic. The distributions are based on an analysis of 33 genes whose expression is male-biased (red), 28 genes whose expression is female-biased (green), and 30 genes with approximately equal expression in adults of both sexes (blue).

all nonsynonymous mutations, however. They include only those that could reach a high enough frequency in a population to have a reasonable chance of being included in a relatively small sample. Excluded from Fig. 1 are what must be a very large number of nonsynonymous mutations whose deleterious effects are so severe that there is essentially no chance of their becoming polymorphic.

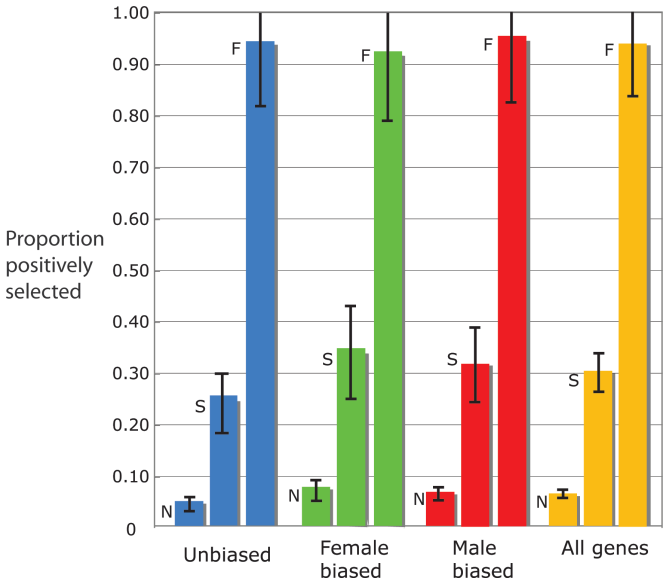

Fig. 2. Estimated proportion of positively selected nonsynonymous mutations among new mutations (N), sample polymorphisms (S), and fixed differences (F) between *D. melanogaster* and *D. simulans*. New mutations include only those that could plausibly become polymorphic or fixed. The error bars are the 95% credible intervals around the means. Blue bars indicate genes expressed approximately equally in adults of both sexes; green bars indicate genes with female-biased expression; red bars indicate genes with male-biased expession; and gold bars indicate all genes combined.

Proportion of Amino Acid Polymorphisms That Are Deleterious. Fig. 2 shows the estimated mean proportion of nonsynonymous substitutions that are positively selected (beneficial). The proportions differ widely among new mutations (N), polymorphisms present in the samples (S), or fixed differences between the species (F). The error bars denote the 95% confidence interval on the estimate of the mean. The results are shown separately for genes with unbiased adult expression (blue), female-biased expression (green), male-biased expression (red), and all genes combined (gold).

For all 91 genes taken together, the fraction of new nonsynonymous mutations that are deleterious averages 0.94 ± 0.01. The preponderance of deleterious new mutations reflects the estimate of $\mu = -5.7 \pm 15.5$ for the average selection coefficient of new mutations across loci.

Our analysis also implies that many of the deleterious nonsynonymous mutations that become polymorphic in the population attain allele frequencies sufficiently high that they account for a significant proportion of the polymorphisms observed in samples. In Fig. 2, among all 91 genes, the expected average proportion of deleterious amino acid polymorphisms in samples is 0.70 ± 0.06. These results again support the widely held belief that most amino acid polymorphisms are deleterious and are maintained in the population by recurrent mutation.

In contrast, while the vast majority of new nonsynonymous mutations and most amino acid polymorphisms are inferred to be deleterious, the model also implies that most amino acid fixations between species are positively selected. In Fig. 2, among all genes taken together, the average proportion of fixed differences that are positively selected is 0.94 ± 0.20.

Weak Positive Selection for Amino Acid Differences Between Species. Although our analysis implies that most amino acid replacements between *D. melanogaster* and *D. simulans* are associated with positive selection, the selection coefficients are very small. The means and standard deviations of the distribution of the scaled selection coefficients of fixed differences for male-biased, female-biased, and unbiased genes are 2.5 ± 0.3, 2.5 ± 0.5, and 2.4 ± 0.4, respectively. These are scaled according to the diploid effective population size, which in the *Drosophila* species considered here is thought to be on the order of 10^6 (28, 29). The unscaled mean selection coefficients among fixed amino acid replacements are therefore on the order of $s = 2.5 \times 10^{-6}$.

Comparison of Genes with Sex-Biased or Unbiased Expression. A previous analysis of these data emphasized evidence for apparently greater selection among genes that are sex-biased in their expression (19). Our model provides a somewhat more nuanced breakdown as to the source of the differences. The comparisons are shown in Table 2, which summarizes the mean values for various features of the data and compares the 33 male-biased genes and the 28 female-biased genes with the 30 unbiased genes. Each P value is based on a null model composed of 10,000 random permutations of the data comparing genes with either male-biased or female-biased expression against genes whose expression is unbiased between the sexes.

Interestingly, the difference between the sex-biased genes and the unbiased genes is not reflected in the proportion of fixed differences that are positively selected ($N_e s > 0$). The differences in the sex-biased genes are mainly in the mean of the mutational distributions and the smaller fraction of slightly deleterious and weakly beneficial mutations that are fixed. For instance, in comparison with unbiased genes, male-biased genes have a significantly higher mean $N_e s$ of the estimated mutational distribution and a significantly lower proportion of nearly neutral ($-1 < N_e s < 1$) fixed differences (Table 2). Furthermore, the male-biased genes have a higher overall fraction of positively selected ($N_e s > 0$) polymorphisms and a greater mean value of

Table 2. Comparison of sex-biased and unbiased genes

Feature	Male-biased expression	Female-biased expression	Unbiased expression
Mean γ of estimated mutational distribution	-5.5 ($P = 0.02$)	-5.4 ($P = 0.02$)	-6.2
Proportion of new mutations with $N_e s > 0$	0.066 ($P = 0.06$)	0.076 ($P = 0.01$)	0.048
Proportion of sample polymorphisms with $N_e s > 0$	0.316 ($P = 0.097$)	0.341 ($P = 0.042$)	0.247
Proportion of fixed differences with $N_e s > 0$	0.952 ($P = 0.34$)	0.919 ($P = 0.77$)	0.940
Mean $N_e s$ of fixed differences	2.6 ($P = 0.07$)	2.5 ($P = 0.22$)	2.4
Mean proportion of fixed mutations with $-1 < N_e s < 1$*	0.209 ($P = 0.02$)	0.210 ($P = 0.03$)	0.239

Each *P* value is based on the results of 10,000 random permutations comparing either male-biased genes or female-biased genes with unbiased genes (genes whose expression does not differ between the sexes).
*Based on sampling from the stationary Markov chain Monte Carlo distribution.

$N_e s$ among fixed differences, although in these comparisons the differences are marginally significant. The female-biased genes show similar patterns when compared with the unbiased genes (Table 2).

Deleterious and Nearly Neutral Amino Acid Replacements. The histograms in Fig. 2 implying a prevalence of positive selection at first seems at odds with the hypothesis that many amino acid replacements fixed between species are nearly neutral (30–32). But the distinction between "near neutrality" and "weak positive selection" is somewhat arbitrary. To approach the issue quantitatively, we estimated the expected proportion of fixed amino acid replacements in which the scaled selection coefficient is smaller than some fixed value of $N_e s$. For each gene the estimated normal density function of the distribution of scaled selection coefficients among new mutations, weighed by the probability of fixation, was numerically integrated from $-\infty$ to $N_e s$ for a fixed value of $N_e s$. The results are shown in Fig. 3 for genes whose expression is male-biased (red), female-biased (green), or unbiased (blue). The proportion of fixed differences that are slightly deleterious ($N_e s < 0$) is by no means negligible. It ranges from 0.02 to 0.11 and across all genes has a mean and 95% confidence interval of 0.05 $\pm$ 0.02. Likewise a significant proportion of fixed differences show weak positive selection ($0 < N_e s < 1$), across all genes averaging 0.17 $\pm$ 0.04 (data not shown).

Across all genes, the average proportion of fixed differences that are positively selected ($N_e s > 0$) is 0.95 $\pm$ 0.02. Positive selection is therefore prevalent. On the other hand, the scaled selection coefficients are very small. For the values of $N_e s$ given in Fig. 3, the means and standard deviations of the estimated proportion of fixed differences that have scaled selection coefficients smaller than $N_e s$ are given by 0.22 $\pm$ 0.05 ($N_e s = 1$), 0.46 $\pm$ 0.08 ($N_e s = 2$), 0.68 $\pm$ 0.07 ($N_e s = 3$), and 0.83 $\pm$ 0.05 ($N_e s = 4$). Because the proportion of amino acid replacements that are nearly neutral depends on what value of $N_e s$ is chosen as an upper limit, one could argue that anywhere from 22% to 83% of fixed amino acid replacements are nearly neutral. This issue is examined further in *Discussion*.

Define $CDF(N_e s)$ as the cumulative density function of fixed amino acid replacements whose scaled selection coefficient is smaller than $N_e s$, and $\alpha(N_e s)$ as the proportion of fixed amino acid replacements whose scaled selection coefficient is greater than $N_e s$. Fig. 4 shows these functions as estimated from the present data. About 50% of all amino acid replacements have $N_e s < 2$, >80% have $N_e s < 4$, and 99% have $N_e s < 7$. Correspondingly, $\alpha(2) = 0.54$, $\alpha(4) = 0.16$, and $\alpha(7) = 0.01$. These estimates contrast with that of $\alpha = 0.25 \pm 0.20$ (6), which assumes three classes of nonsynonymous mutations (deleterious, neutral, and beneficial) with deleterious mutations being so deleterious and beneficial mutations being so beneficial that

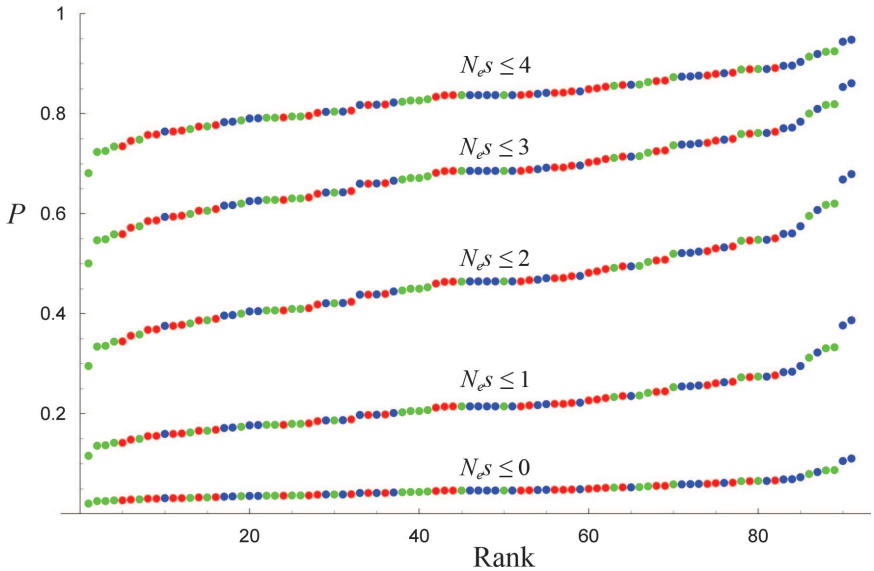

Fig. 3. Estimated proportion of fixed amino acid replacements between *D. melanogaster* and *D. simulans* whose scaled selection coefficient is less than various specified valus of $N_e s$, ordered by rank among all 91 genes. N_e is the diploid effective population number and s is the conventional selection coefficient. Blue dots indicate genes expressed approximately equally in adults of both sexes; green dots indicate genes with female-biased expression; and red dots indicate genes with male-biased expression.

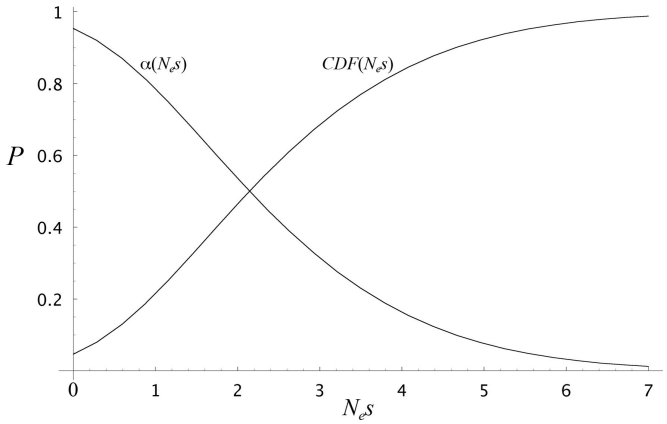

Fig. 4. Inferred cumulative density function (*CDF*) of the scaled selection coefficients among fixed amino acid replacements in 91 genes between *D. melanogaster* and *D. simulans*. N_e is the diploid effective population number, and s is the conventional selection coefficient. $CDF(N_es)$ is the average proportion of amino acid differences whose scaled selection coefficient is smaller than N_es, and $\alpha(N_es)$ is the proportion of amino acid differences whose scaled selection coefficient is greater than N_es.

neither class contributes significantly to polymorphism. Our model takes slightly deleterious and weakly beneficial mutations into account, and as shown in Fig. 2 these classes of mutations do contribute substantially to amino acid polymorphisms. The estimate $\alpha = 0.25$ corresponds roughly to $\alpha(1.15)$ in Fig. 4, hence it implies a threshold for near-neutral effects of $N_es = 1.15$.

Discussion

Our model makes a number of assumptions that should be emphasized. The theory assumes mutation-selection-drift equilibrium, invokes diffusion theory, stipulates independence between nucleotide sites, and posits additivity of fitness effects of mutations at different nucleotide sites. The first assumption could be undermined by demographic factors such as population bottlenecks or expansions, the second could be compromised by very strong positive selection, the third may be challenged for genes in regions of the genome with reduced recombination, and the fourth could be subverted by potential epistatic effects of nonsynonymous mutations in the same gene (9). Additional study is needed to determine how robust the model may be to small departures from the assumptions.

Several features of the results give some reassurance because they support plausible notions and other evidence that most nonsynonymous mutations and many nonsynonymous polymorphisms are deleterious (1, 5, 8–10). Our analysis implies that some 19 of 20 new amino acid replacements are deleterious with an average fitness reduction on the order of five times the reciprocal of the effective population size. These estimates pertain only to the subset of nonsynonymous mutations whose effect are not so severe as to preclude their becoming polymorphic, but they support other evidence that selection against deleterious mutations plays in key role in shaping patterns of genetic variation in *Drosophila* (33). Likewise, we estimate that ≈7 of 10 amino acid replacements that are polymorphic in samples are deleterious.

One feature of our results that might animate some surprise is the high proportion of amino acid fixations between species that show positive selection, ≈95% in our data. This finding seems to reflect what Wallace (34) called the "overwhelming odds against the less fit." It can be appreciated quantitatively by noting that a new mutation with $N_es = 2$ is eight times more likely to be fixed than one with $N_es = 0$ and ≈3,000 times more likely to be fixed than one with $N_es = -2$. There would be a preponderance of deleterious fixations if beneficial mutations were vanishingly rare. But for mutations with selective effects near neutrality, Fisher (35) argued from analogy that the proportion of beneficial mutations should actually be close to one-half:

> "The conformity of these statistical requirements with common experience will be perceived by comparison with the mechanical adaptation of an instrument, such as the microscope, when adjusted for distinct vision. If we imagine a derangement of the system by moving a little each of the lenses, either longitudinally or transversely, or by twisting through an angle, by altering the refractive index and transparency of the different components, or the curvature, or the polish of the interfaces, it is sufficiently obvious that any large derangement will have a very small probability of improving the adjustment, while in the case of alterations much less than the smallest of those intentionally effected by the maker or the operator, the chance of improvement should be almost exactly half."

This inference also follows from the assumption of a normal distribution of selection coefficients, because adjacent small intervals of the same width on opposite sides of $N_es = 0$ will have approximately equal areas.

The results in Fig. 4 might give satisfaction to both selectionists and nearly neutralists. On the one hand, ≈95% of the fixed amino acid replacements are positively selected; on the other hand, most of the selection coefficients are small (average $N_es \approx 2.5$). As emphasized by Nei (32), the fate of mutations with such a small selective advantage will be determined in large part by random genetic drift. Nevertheless when a large number of sites are examined (>58,000 nonsynonymous sites in the present case), the statistical signal of weak positive selection is evident. These results suggest that, across the genome as a whole, weak positive selection plays an important role in the evolution of protein sequences.

What fraction of amino acid replacements should be considered as nearly neutral is a matter of definition. Ohta (36) has stressed that the key feature of nearly neutral mutations is that their fate in the population depends on both selection and random genetic drift and has suggested that an absolute value of $N_es < 2$ would be suitable as a definition. For our data, this threshold implies that ≈46% of fixed amino acid replacements are selectively nearly neutral. One might also regard a mutation as selectively nearly neutral if its probability of fixation were <10 times that of a truly neutral allele; with this definition $N_es = 2.5$ and the proportion of fixed amino acid replacements that are selectively nearly neutral is 58%. A threshold of $N_es = 4$ yields a proportion of selectively neutral amino acid fixations of 0.84. Nei (32) has given reasons a much larger threshold could be defended. If $N_es = 7$ the proportion of nearly neutral amino acid replacements becomes 0.9878, and for $N_es = 10$ it becomes 0.9996. Our model also explicitly assumes that all synonymous polymorphisms and replacements are neutral or nearly neutral. Because our model as presently formulated pertains only to coding regions, which are very sparse in complex genomes such as the human genome, the model is uninformative with regard to the selective effects of mutations in introns and other noncoding regions.

What might be the molecular mechanism behind extremely small selective effects of amino acid replacements? There is no definitive evidence, but DePristo *et al.* (37) have suggested a model of protein evolution in which many amino acid replacements result in very small differences in protein stability, aggregation, or degradation. Their model is based on the observation that many native proteins have a free energy of folding equivalent to only a few hydrogen

bonds. Most amino acid replacements are assumed to be approximately additive with respect to their effects on stability, aggregation, or degradation, and within broad limits are selectively nearly neutral. Outside these limits increased instability results in greater aggregation and degradation and a lower equilibrium concentration of active protein, whereas increased stability results in resistance to degradation and a greater concentration of active protein. The effect of any amino acid replacement therefore depends on its context. What is slightly deleterious in one genetic background may be mildly beneficial in another. However, most amino acid replacements with small effects are expected to be deleterious. It has been noted that the low frequencies of most amino acid polymorphisms in natural populations of *E. coli* and *Salmonella enterica* imply that the mutations are slightly deleterious (38), and in the context of the stability-aggregation-degradation model it is of interest that virtually all of these are physically located in regions of high solvent accessibility on the "outside" of the molecule (39).

The mapping of stability and aggregation onto fitness implies that amino acid replacements would show epistasis at the level of fitness even though they may be additive in their contribution to the free energy of folding. The model of selection presented here does not capture these interaction effects on fitness, nor does it capture the potential context dependence of amino acid replacements. Any model that takes such interactions into account might have to be quite protein-specific. Our model is more generic and may instead be thought of as estimating the "effective" selection among nonsynonymous mutations in a set of ideal loci in which all nucleotide sites are independent and all selective effects constant and additive, and whose levels of polymorphism and divergence are similar to those observed among the actual loci.

Methods

In principle, after initialization, each each step of the Monte Carlo Markov chain in the $(3k + 4)$-dimensional parameter space of vectors $(\gamma_i, \theta_{s,j}, \theta_{r,j}, T, \sigma_w, \mu, \sigma_b)$ could be composed of a series of Metropolis-random-walk (40) or Gibbs-sampler (41) substeps, with each substep updating a single 1D or 2D component of the vector of parameters. The structure of the model is such that $\theta_{s,j}$ and $\theta_{r,j}$ have Gibbs-sampler updates based on gamma distributions, and (μ, σ_b) together can be updated by using a 2D inverse-gamma-normal Gibbs update (27). The other components would be updated by using Metropolis random-walk steps. Updating σ_w is the most time-consuming step in this algorithm because each update requires the numerical calculation of up to $4k$ double integrals.

In practice, the method described above took extremely long to converge, with some data sets converging to different distributions depending on the initial point. The reason was that updates of (μ, σ_w) in particular, and to a lesser extent $(\mu, \sigma_w, \sigma_b, \gamma_i)$, were highly autocorrelated. What was done was to use a long run of the process described above to estimate a joint covariance matrix for $(\mu, \sigma_w, \sigma_b)$. The Metropolis update for σ_w was then replaced by a joint Metropolis update of $(\mu, \sigma_w, \sigma_b)$ based on a 3D normal distribution with a larger step. A linear or skew transformation of the γ_i was made at the same time corresponding to the change in (μ, σ_b). The resulting $(k + 3)$-dimensional update is not of Metropolis–Hastings form because it is defined by a singular motion in $(k + 3)$ dimensions, but it does satisfy the detailed balance condition (42, 43) and hence preserves the posterior likelihood. The resulting process converged to the same distribution independent of starting position. Trace plots of the hyperparameters $(\mu, \sigma_w, \sigma_b)$ appeared highly random.

The results are based on 10 consecutive subchains of 200,000 samples each after a burn-in of 1 million iterations. Samples were taken every 10 iterations to reduce autocorrelation, so there was a total of 21,000,000 iterations. Acceptance proportions for the Metropolis random-walk component updates ranged from 0.17 to 0.32. The Gelman *et al.* (27) statistic for convergence ranged from 1.000 to 1.020.

We thank Tomoko Ohta and Masatoshi Nei for their careful reading and helpful comments on the manuscript. This work was supported by National Institutes of Health Grants GM68465 and GM61351 (to D.L.H.), National Science Foundation Grant DMS-0107420 (to S.A.S.), and Deutsche Forschungsgemeinschaft Grant PA 903/2 (to J.P.).

1. Sawyer SA, Dykhuizen DE, Hartl DL (1987) *Proc Natl Acad Sci USA* 84:6225–6228.
2. McDonald JH, Kreitman M (1991) *Nature* 351:652–654.
3. Sawyer SA, Hartl DL (1992) *Genetics* 132:1161–1176.
4. Smith NGC, Eyre-Walker A (2002) *Nature* 415:1022–1024.
5. Fay JC, Wyckoff GJ, Wu CI (2002) *Nature* 415:1024–1026.
6. Bierne N, Eyre-Walker A (2004) *Mol Biol Evol* 21:1350–1360.
7. Shapiro JA, Huang W, Zhang C, Hubisz MJ, Lu J, Turissini DA, Fang S, Wang H-Y, Hudson RR, Nielsen R, *et al.* (2007) *Proc Natl Acad Sci USA* 104:2271–2276.
8. Fay JC, Wyckoff GJ, Wu C-I (2001) *Genetics* 158:1227–1234.
9. Sawyer SA, Kulathinal R, Bustamante CD, Hartl DL (2003) *J Mol Evol* 57:S154–S164.
10. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, *et al.* (2006) *Nature* 437:1153–1157.
11. Mustonen V, Lässig M (2007) *Proc Natl Acad Sci USA* 104:2277–2282.
12. Bustamante C, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL (2002) *Nature* 416:531–534.
13. Whittam TS, Nei M (1991) *Nature* 354:115–116.
14. Hartl DL, Clark AG (2007) *Principles of Population Genetics* (Sinauer, Sunderland, MA).
15. Eyre-Walker A (2002) *Genetics* 162:2017–2024.
16. David JR, Capy P (1988) *Trends Genet* 4:106–111.
17. Lachaise D, Cariou M, David JR, Lemeunier F, Tsacas L, Ashburner M (1988) in *Evolutionary Biology*, eds Hecht MK, Wallace B, Prance GT (Plenum, New York), Vol 22, pp 159–227.
18. Li H, Stephan W (2006) *PLoS Genet* 2:e166.
19. Pröschel M, Zhang Z, Parsch J (2006) *Genetics* 174:893–900.
20. Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D (2003) *Genetics* 165:1269–1278.
21. Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, Andrews J, Eastman S, Oliver B (2003) *Science* 299:697–700.
22. Ranz JM, Castillo-Davis CI, Meiklejohn CD, Hartl DL (2003) *Science* 300:1742–1745.
23. Gibson G, Riley-Berger R, Harshman L, Kopp A, Vacha S, Nuzhdin S, Wayne M (2004) *Genetics* 167:1791–1799.
24. Meiklejohn CD, Kim Y, Hartl DL, Parsch J (2004) *Genetics* 168:265–279.
25. Wright S (1938) *Proc Natl Acad Sci USA* 24:253–259.
26. Gilks R, Richardson S, Spiegelhalter DJ (1996) *Markov Chain Monte Carlo in Practice* (Chapman & Hall, London).
27. Gelman A, Carlin JS, Stern HS, Rubin DB (2003) *Bayesian Data Analysis* (CRC, Boca Raton, FL).
28. Akashi H (1995) *Genetics* 139:1067–1076.
29. Akashi H (1996) *Genetics* 144:1297–1307.
30. Ohta T (1973) *Nature* 246:96–98.
31. Ohta T (1992) *Annu Rev Ecol Syst* 23:263–286.
32. Nei M (2005) *Mol Biol Evol* 22:2318–2342.
33. Haag-Liautard C, Dorris M, Maside XR, Macaskill S, Halligan DL, Charlesworth B, Keightley PD (2007) *Nature* 445:82–85.
34. Wallace AR (1892) *Nat Sci* 1:749–750.
35. Fisher RA (1930) *The Genetical Theory of Natural Selection* (Oxford Univ Press, Oxford).
36. Ohta T (2002) *Proc Natl Acad Sci USA* 99:16134–16137.
37. DePristo MA, Weinreich DM, Hartl DL (2005) *Nat Rev Genet* 6:678–687.
38. Hartl DL, Boyd EF, Bustamante CD, Sawyer SA (2000) in *Genomics and Proteomics*, ed Suhai S (Plenum, New York), pp 37–49.
39. Bustamante CD, Townsend JP, Hartl DL (2000) *Mol Biol Evol* 17:301–308.
40. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) *J Chem Phys* 21:1087–1091.
41. Geman S, Geman D (1984) *IEEE Trans Pattern Anal Machine Intelligence* 6:721–741.
42. Chib S, Greenberg E (1995) *Am Stat* 49:327–335.
43. Liu JS (2001) *Monte Carlo Strategies in Scientific Computing* (Springer, New York).